

# Chapter 7

## Situational Judgment Tests and Their Use for Teacher Selection



**Abstract** In this chapter we will sharpen our focus to look closely at situational judgment tests (SJTs), typically used for large-scale screening of applicants to training programs. Although SJTs have a solid research foundation and are commonly used for selection into training and employment in diverse professional fields, especially health-related fields, they have rarely been implemented in teacher education. In this chapter we look at the research and theory behind the use of SJTs for selecting the best possible teachers and provide some direction for how this method can be developed for use by ITE programs and other education organizations.

In Chap. 6, we took a historical view of teacher selection methods, and also examined methods that are currently implemented around the world. In this chapter we will sharpen our focus to look closely at situational judgment tests (SJTs), typically used for large-scale screening of applicants to training programs. Although SJTs have a solid research foundation and are commonly used for selection into training and employment in diverse professional fields, especially health-related fields, they have rarely been implemented in teacher education. In this chapter we look at the research and theory behind the use of SJTs for selecting the best possible teachers and provide some direction for how this method can be developed for use by ITE programs and other education organizations.

### 7.1 Situational Judgment Tests

SJTs have become increasingly popular in the last two decades because they show higher levels of predictive validity than other screening measures assessing non-cognitive attributes, are easy to use, and are well-received by applicants (e.g., Klassen et al., 2014). SJTs are a measurement method well suited for measuring judgment in challenging situations, and usually consist of a ‘stem’ and a series of response options.

The stem consists of a range of contextualized work-related scenarios presented in text or video format, and the response options provide a list of possible courses of action, usually preceded with the phrase, ‘What *should* you do?’ (procedural knowledge) or ‘What *would* you do?’ (behavioral intentions). Early versions of SJTs were created by military psychologists to select soldiers to join the armed forces in World War Two where the tests included a series of detailed and realistic scenarios that described challenging situations likely to be encountered in military settings. The early military SJTs were useful in several ways: they gave potential candidates a taste of what life as a soldier might hold, and they gave recruiters insight in how judgment was displayed by potential officers (Lievens & De Soete, 2015).

**Theory underpinning SJTs.** The theory underpinning SJTs—implicit trait policy—refers to an individual’s implicit beliefs about the effectiveness of expressing particular personality traits in particular situations (Whetzel & McDaniel, 2009). SJTs do not explicitly measure personality or other non-cognitive attributes but are designed to capture useful information about personality traits *indirectly* by asking people to evaluate work-related scenarios and then to judge the effectiveness of response options. Some additional theoretical foundations for SJTs can be located in Sternberg’s theory of successful intelligence, whereby procedural knowledge in complex situations is often tacit (e.g., Elliott et al., 2011), and in Boyatzis and Kelner’s (2010) theory of links between attributes as the behavioral manifestation of implicit motives. However, the theory of *implicit trait policy* is most often used to describe the framework supporting the use of SJTs.

SJTs are considered a measurement method, and as such can be designed to capture a range of non-cognitive attributes, including Big Five personality traits such as agreeableness, extraversion, and conscientiousness (e.g., Hooper et al., 2004). SJTs can be designed to measure other related non-cognitive domains—including motivation, resilience, professional integrity, and empathy—derived from a careful job analysis of workplace demands (e.g., Patterson, Lievens, et al., 2013). A growing body of research shows that SJTs are a reliable and valid approach to make selection decisions in a range of professional contexts (Whetzel et al., 2020).

**Context-specific or context-general SJTs?** SJTs can be constructed to reflect a particular context, for example, scenarios representing a school environment for teacher selection SJTs, or to reflect more general situations that one might encounter in daily life. There is divided opinion on the importance of contextualization of SJTs. Lievens and Motowidlo (2016) argued that SJTs tap general domain knowledge that requires an understanding of the utility of expressing certain traits across a range of work settings. In such a model, SJTs are designed to measure specific constructs in a clear and explicit way, without reference to a particular situation. Other SJT theorists and researchers disagree. Harris et al. (2016) countered the ‘situation-free’ SJT approach by noting that general domain knowledge is always contingent on the use of contextual and situational cues. In trait activation theory (TAT), it is the interaction between person and situation that explains behaviour; a trait will only be expressed when a trait-relevant situation demands the activation of that particular trait in that particular situation (Harris et al., 2016).

In a similar fashion, Bandura's social cognitive theory proposes that personal characteristics are not formed and expressed in isolation; rather, it is the reciprocal interaction between personal characteristics, behaviour, and the environment that forms the basis of human agency (Bandura, 1999). Fan et al. (2016) argue that although general domain knowledge is an important feature of SJTs, it is the ability to understand when and how to express certain traits that separates SJTs from disembodied measures of personality or other interpersonal attributes. A series of studies recently published by Freudenstein et al. (2020) tested the importance of situation construal in SJTs. The authors found that test-takers' perceptions of the situation predicted responses even after controlling for personality, emotion recognition, and mental ability, and that situation construal plays a pivotal role in determining SJT responses. For SJTs used in teacher selection, providing job-relevant situational cues may be essential to understand how particular attributes are activated in authentic classroom environments.

**Research on the use of SJTs for selection.** The use of SJTs as an alternative to conventional selection tests for entry into professional training has received considerable recent research attention. The surge in interest is due to the effectiveness of SJTs for predicting job performance (e.g., Christian et al., 2010): SJTs have been shown to be better predictors of job performance than conventional personality tests (e.g., Shultz & Zedeck, 2012), and when tailored to specific contexts, are useful for selection purposes in a wide range of fields (Patterson et al., 2015). In addition, SJTs tend to display stronger face and content validity than conventional non-academic measures due to their close correspondence to the work-related situations that they describe (Whetzel & McDaniel, 2009). In addition, SJTs constructed by researchers working in collaboration with expert practitioners are less susceptible to coaching effects and faking than other kinds of selection tests (e.g., conventional personality tests).

Recent empirical studies and meta-analyses show that SJTs administered as selection tools at the beginning of training programs can be reliable and robust predictors of subsequent job performance (Patterson, Lievens, et al., 2013). SJTs have been used for selection into training programmes in a range of professions, including dentistry, law and medicine (see Chap. 5 for more detail). In medicine, SJTs have been successfully validated for use in selection into foundation year training in the UK (Patterson, Tavabie, et al., 2013) and are widely used across medical schools for selection purposes in the UK. In the United States, Shultz and Zedeck (2012) reported that SJTs were a better predictor of lawyer effectiveness than the conventional tests used for selection into highly competitive law schools, and furthermore, were less prone to inter-group differences (i.e., gender, SES, and ethnicity) than conventionally-used selection metrics (i.e., Law School Admissions Test and grade point average). SJTs show less inter-group bias than other selection methods such as tests of cognitive ability and interviews and are perceived to be fair by candidates (Patterson et al., 2015). In Table 7.1, we present a brief summary of research on SJTs in diverse professional fields, along with their validity evidence.

**Table 7.1** SJTs Used for selection in various disciplines

Study	Context	Validity Evidence
Lievens and Sackett (2012)	Admission into <b>medical school</b>	SJTs (procedural knowledge about interpersonal behavior) predicted internship performance and job performance ( $r = 0.21$ ) 9 years after SJT administration
Koczwara et al. (2012)	Admission into <b>advanced medical training</b>	SJTs were the best single predictor of performance in selection center
Patterson et al. (2012)	Admissions into <b>advanced dentistry training</b>	SJTs showed significant correlations ( $r = 0.43$ ) with entrance interview (concurrent validity)
Shultz et al. (2012)	Developing new measures for <b>law school admissions</b>	SJTs showed significant correlations with 23 of 26 lawyering effectiveness factors
Bateson et al. (2014)	Selection of <b>service employees</b>	SJTs showed predictive validity for selection of service employees at the start of the recruitment process
Klassen et al. (2020)	Admissions into teacher education	SJTs used for selection showed predictive validity for performance in teaching placements ( $r_s 0.24-0.30$ , $p < 0.01$ )

**Reliability and predictive validity of SJTs.** Most SJTs tend to have lower internal consistency than other tests measuring non-cognitive attributes, largely due to their multidimensional nature, with a review of SJT reliability showing a weighted corrected coefficient  $\alpha$  of 0.46 (Catano et al., 2012). Kasten and Freund (2016) found that SJT internal consistency was higher for low stakes tests, for tests using theoretical, versus expert-based or empirical scoring, and for tests using Likert-type response scales, rather than ‘pick-best’ scoring approaches. The authentic situations on which SJTs are built tend to be complex, ‘messy’, and reflective of more than one non-cognitive attribute, even when an individual scenario is designed to represent a single attribute.

SJTs tend to be constructed to cover multiple domains, explaining their sometimes-lower internal consistency (but higher predictive validity) than single construct measures (e.g., cognitive ability or personality). However, reliability is dependent on test length and item heterogeneity. Internal consistency (i.e., alpha) may not be the best reliability index if item heterogeneity is high in the SJT (Catano et al., 2012), and some form of test–retest reliability or split-half estimates (e.g., Whetzel et al., 2020) will supplement reliability estimations. Internal consistency of teacher selection SJTs using a rating approach was shown to be acceptable (e.g.,  $\alpha$

= 0.78 in Klassen et al., 2020;  $\alpha = 0.79$  in Klassen et al., 2017; and  $\alpha = 0.70$  in brief video format, Bardach et al., 2020).

Factor analysis of SJT content typically results in ambiguous factor structures unless cross-loadings are allowed. Factor analytic approaches need to explicitly model the multidimensionality of SJTs at the item level, not just for the test as a whole. The development strategy for SJTs represents a trade-off in assessment of non-cognitive attributes. An inductive or ‘bottom-up’ approach (e.g., using critical incidents) may result in a more predictive test, but one with lower internal consistency, whereas an SJT developed using a deductive, ‘top-down’ approach targeting specific constructs may result in a more internally consistent measure.

Meta-analytic research indicates that SJTs generally have good predictive validity (corrected  $r = 0.34$ ; McDaniel et al. 2001). What has been difficult with SJTs in general has been establishing exactly which constructs are being measured and they are sometimes criticized as a ‘black box’ measurement method. A study investigated whether SJTs predicted job performance above and beyond cognitive ability, job experience, job knowledge, and conscientiousness in three samples (Clevenger et al., 2001), with SJTs predicting job performance in all three samples.

**Scoring options for SJTs.** Scoring for SJTs involves consideration of several options. Typically, SJTs are scored by comparing applicants’ judgments with the judgments expressed by subject matter experts (SMEs). The judgment tasks in an SJT are designed to assess contextualized judgment and are based on the notion that situation-specific judgments and responses reflect implicit personality traits that have a causal effect on job performance. In contrast, conventional personality tests ask individuals to describe themselves directly, opening up the likelihood that candidates will choose responses that portray their personality in the best possible light, but possibly inaccurately. A second scoring option is empirical scoring, where scoring is determined either by consensus (i.e., ‘crowd wisdom’) or by examining applicant data and specifically the correlations of each response option to a criterion score (e.g., performance score on an important criterion). A third option, especially for SJTs that are more purely ‘construct-driven’ is to use a theoretically derived scoring key whereby the scoring pattern is determined by reference to, and interpretation of, the construct underlying the SJT content (e.g., Tiffin et al., 2020).

**Traditional vs. construct driven SJTs.** The traditional approach to developing SJT content is a ‘bottom-up’ or inductive approach, where SMEs (usually led by a psychologist or consultant) gather together to elicit ‘critical incidents’ (*Think of a time when a new trainee faced a challenging classroom situation requiring careful judgment. What did this trainee do? Was the response appropriate in your view? What other options might s/he have considered?*). Responses to these critical incident questions form the basis of SJT scenarios, with content reflecting authentic workplace situations and challenges. Following this path of test development typically leads to an SJT that is contextualized, accepted by candidates, predictive of the job it is built around, but hampered by conceptual and psychometric issues, and in particular, lack of a clear factor structure. There are clear benefits to developing SJTs using a traditional inductive approach, but some disadvantages, and in particular, problems with relating the content to specific domains and constructs.

In contrast, developing a construct-driven SJT depends on a ‘top-down’ approach, typically led by a psychologist or team of psychologists who focus more on a single target trait (e.g., integrity), and less on a particular context. The scenario is designed to elicit a particular trait, and the response options represent degrees of the target trait. Reliability coefficients tend to be higher for single-construct SJTs, and test scores tend to correlate more highly with personality measures. Predictive validity of construct driven SJTs is still emerging, although tests developed using this approach tend to show high correlations with other measures of the target trait. A multimedia SJT assessing emotion management showed similar validity to conventional measures of emotion management, although the test was not used to predict workplace performance. Construct-driven SJTs may be more prone to faking, similar to the case with conventional personality tests, where the appropriate course of action may be easier to detect when a series of scenarios all focus on the same construct with a detectable theme (Tiffin et al., 2020). Some promising work on construct driven SJTs relevant to teacher selection is currently being conducted by Bostwick and Durksen at the University of New South Wales in Australia in collaboration with the Teacher Selection Project (K. Bostwick, personal communication, October 2020), with development of a prototype SJT assessing growth mindset in teachers recently piloted.

Recent studies (e.g., Klassen et al., 2020) have used an integrated ‘construct-informed’ approach, in which target attributes are developed before the scenario development process, and critical incidents are mapped onto these attributes. A combined inductive-deductive approach allows for a priori non-cognitive attributes to be identified early on in the test development process (i.e., using a deductive approach), while allowing for ‘bottom-up’ or inductively derived attributes to emerge during the development process. Using an integrated construct-informed approach leads to building the scenarios and response options of the SJT on a foundation of identifiable domains, although factor analyses typically show the existence of single overarching factors representing judgment about effective behaviors in particular contexts (Patterson et al., 2015).

**Applicant reactions to SJTs.** Research on applicant reactions to selection processes has been based on evaluation of *procedural justice*, or the perceived fairness of the methods used for decision making, and *distributive justice*, or the fairness of the outcome of the selection process (Patterson et al., 2011). Measures of non-cognitive attributes such as interpersonal skills and empathy may not be viewed as positively as ‘fact-based’ assessments; indeed, Patterson et al. (2011) found that for selection into specialist medical training, a clinical skills test was more favourably received by candidates than an SJT assessing non-clinical judgment, even though the SJT was a better predictor of subsequent outcomes. Most studies show that SJTs are favourably received by applicants, and that video-based SJTs are preferred to text-based SJTs. Bardach et al. (2020) found that video-based SJTs were more engaging than text-based SJTs, but applicants did not rate the video format as fairer or more job-related than the text version. Klassen et al. (2014) evaluated applicant reactions to an SJT for admission into ITE using a mixed methods approach and found that applicants were generally favorable about the relevance of the SJT content and the

appropriateness of its use for selection, but raised questions about procedural justice issues related to (in)experience: “It’s hard to judge (the scenarios) if you’ve never been in a certain situation before”, and that the SJTs measured skills that “should be taught during teacher training” (p. 116). However other researchers have shown that SJTs that are contextualized and show higher fidelity to the job are preferred by candidates due to their relevance and the realistic nature of the scenarios (Whetzel & McDaniel, 2009).

## 7.2 Situational Judgment Tests for Teacher Selection

SJTs have rarely been implemented for selection into teacher education or for entry into the profession, but their use is increasing in the last decade. In a review of teacher selection methods (Klassen & Kim, 2019), the authors found that it was rare for teacher selection programs to be built on methods with a published evidence base (i.e., using SJTs, MMIs, or other evidence-based methods). Research on using SJTs for selecting teachers is slowly emerging, with most studies in the last decade coming from the Teacher Selection Project group in the UK (<https://www.teacherselect.org/>), and some from the TCAT group in Australia (e.g., Bowles et al., 2014). In 2014, the Teacher Selection Project group published one of the first articles on teacher selection using SJTs (Klassen et al., 2014), describing applicant reactions to SJTs used in parallel to other selection methods (see Fig. 7.1 for a sample teacher selection SJT). Results showed that SJTs administered to primary and secondary teaching applicants were generally well received, with primary applicants expressing more favorable opinions of the test than secondary applicants.

The research base investigating the use of SJTs to predict teaching performance has grown at pace in the last 10 years, including research on the predictive validities

**Situational Judgment Test Sample Item**

You are walking into school when the parent of one of your students, Mr Andrews, asks if he can speak to you. He informs you that his son, Callum, was hit on the playground at lunchtime yesterday by another student, Jack, and came home very upset. You have been unaware of the incident until now, and you don't know whether similar incidents have occurred in the past.

*Rate the appropriateness of each of the options in terms of what you should do as a first-year teacher (Inappropriate to Appropriate)*

- Look in both of the students' school records to establish if similar incidents have occurred before
- Reassure Mr. Andrews that the incident will be investigated
- Ask Jack's parents to come in for a meeting to discuss Jack's behavior

**Fig. 7.1** Example text-based SJT from the Teacher Selection Project

of SJTs and other selection methods in a range of settings. In Klassen et al., 2020, a 25-item SJT was administered online along with two competency-based essays in order to screen candidates for invitation to an ‘assessment center interview day’. The assessment center included a 1–1 interview, a ‘case study’ worked on in small groups, and a teaching demonstration. The SJT was correlated  $r = 0.42$ ,  $p < 0.01$  with the teaching demonstration and  $r = 0.46$ ,  $p < 0.01$  with the assessment center total (but not with the group case study). The other screening measures (i.e., the competency-based essays) were not significantly correlated with the teaching demonstration ( $r_s = 0.14$  to  $0.20$ ,  $p_s = ns$ ) but were correlated with overall assessment center performance ( $r_s = 0.24$ ,  $0.22$ ). Hierarchical regression showed that scores on the SJT contributed unique variance to the prediction of assessment center performance. The conclusion of this study was that SJTs were a useful screening method—predictive and efficient—in comparison to the other methods used.

A recent study explored how SJTs used for selection into primary and secondary ITE programs predicted teaching performance during a major teaching placement approximately six months after selection (Klassen & Rushby, 2019). The study showed that the ‘conventional’ methods employed during the selection process—math and English tests, a group problem-solving task, and a 1–1 interview—were not significant predictors of teaching performance on the teaching placement six months after selection (primary program,  $r_s = -0.13$  to  $0.09$ ; secondary program  $r_s = -0.14$  to  $0.15$ ). In contrast, SJTs were significantly predictive of teaching performance both at the primary level ( $r = 0.24$ ,  $p < 0.05$ ) and secondary level ( $r = 0.35$ ,  $p < 0.05$ ). The results from this study showed that an SJT tailored for teacher selection purposes may provide an effective way to systematically evaluate large numbers of applicants to ITE programs.

**Video SJTs.** Video formats of SJTs are intuitively appealing, with the opportunity to provide applicants with engaging animations or live action (with human actors) as they work through an SJT. Video SJTs, whether using animation or live actors, provide a higher level of realism and allow the test designer to add details (e.g., facial expressions and body language) not easily represented in text. However, video SJTs are expensive to produce, and revisions made to item content can be time-consuming and costly. Recent research has compared video- and text-based SJTs in a range of settings. Lievens and Sackett (2006) compared the predictive validity of video and text SJTs and found that video-based SJTs measuring interpersonal skills had significantly higher predictive and incremental validity than the text SJTs using the same content. The authors suggested that the video format provided extra sources of information, leading to higher accuracy and fidelity. In addition, the text-based SJTs were correlated more strongly with cognitive predictors than were the video SJTs, suggesting that the video format may be a better one for use in assessing non-cognitive variables.

Bardach et al. (2020) from the Teacher Selection Project group recently compared video and text formats for a teacher selection SJT. Prospective teachers were randomly assigned to one of three SJT conditions: 3D animated video with accompanying text, 3D animated video without text, and text only (see Fig. 7.2 for example image). The authors examined how the format of SJTs would be associated with



	Inappropriate	Somewhat inappropriate	Somewhat appropriate	Appropriate
Go to get assistance from a senior member of staff	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Ignore that Mark is using his phone and continue with the lesson	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Ask Mark to put the phone away again, but more sternly	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Give school sanctions for having a mobile phone in class and for refusing the teacher's instruction	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Fig. 7.2 Example video-based SJT from the Teacher Selection Project

applicant reactions and subgroup (gender and ethnicity) differences. No differences in scores between the three formats were found, but participants found the two video conditions more engaging than the text format. Females scored significantly higher than males in the text format SJT, consistent with much of the SJT literature, but that difference disappeared in the video format SJTs, consistent with the findings by Bruk-Lee et al. (2016). Ethnicity effects (participants from majority groups scoring higher than minority groups) were consistent in all three SJT formats, although the mean score differences were not large, typically about 3 points (e.g., for video with text condition,  $M_{majority} = 146.4, SD = 6.2, M_{minority} = 143.5, SD = 4.8$ ). The study concluded that the benefits of video SJTs pertaining to applicant engagement and reduced gender effects should be weighed against the resources (time, money, expertise) needed to produce SJTs in this format. In addition, further exploration was needed to understand the persistence of ethnic group scoring differences.


### 7.3 Developing SJTs: A Collaborative Approach

Developing SJTs involves creating scenarios and writing response options, with content generated through a collaboration between test-developers and ‘subject matter experts’, or SMEs. Research has shown that a collaborative approach to item-writing improves the conceptual and psychometric characteristics of test items, and in particular, the item discrimination and reliability of items (Abozaid et al., 2017). For ‘traditional’ SJTs (i.e., measuring overall situational judgment rather than specific constructs), development typically involves creating scenarios using a ‘critical incident’ approach, where SMEs (including job incumbents, supervisors, trainees) describe past experiences of specific problems faced during work, the action taken or considered to address the problem, and the outcome of the actions. The construction of each scenario is thus highly contextualized and authentic, ensuring fidelity with the actual workplace. For teacher selection SJTs, SMEs might include experienced teachers, teacher educators, and school leaders who have experience in working with novice teachers. The authenticity of SJTs depends on the ‘real-life’ experiences of those involved in developing the test content.

One approach to building content for SJTs is to use a ‘workshop approach’ in which test developers and SMEs work collaboratively to develop test content. The workshops consist of a gathering of experienced educators, primarily teachers, principals, and teacher educators who are brought together to determine target attributes, and to develop and test content. An important consideration when setting up the workshops is to ensure that the SMEs invited to participate reflect the diversity of the potential applicants; teacher workforces are frequently unrepresentative of the general population of the students they teach (Carter Andrews et al., 2019). Inviting a diverse SME group helps ensure that the content of the selection methods represents the target population, and by extension, the teaching workforce. Although organizational psychologists frequently develop SJT content through paired writing where SMEs collaborate with item-writers in dyads, a workshop with multiple participants and open communication can deliver higher quality content—at least at the early stages—for complex, multi-faceted professions like teaching, where contextual differences are marked. The development of SJTs for teacher selection is carried out in three phases—identifying target attributes, creating content, and pilot-testing and administration—with eight steps (see Fig. 7.3 for the proposed framework). Each of the steps is carried out through close collaboration between education experts and the test developers.

**Step 1: Test specification.** The first step in developing an SJT for teacher selection is to specify the purpose of the test, the feasible length of the test, the item types and response formats, how the test will be administered, and the kinds of information that need to be generated by test administration. Important considerations at this step are delivery method (online, paper-and-pencil, proctored, unsecured at home), and response formats (rating, ranking, best and worst, etc.). Delivery methods have become more streamlined in recent years, and online test delivery, whether the test is administered on-site or remotely, is infinitely more desirable for reasons of data

	Identifying target attributes		Creating content			Pilot testing and administration		
Task	Step 1: Test specification	Step 2: Select and define key attributes	Step 3: Review of existing items in target context	Step 4: Generate new items with review	Step 5: Determine scoring key.	Step 6: Design and pilot SJT	Step 7: Psychometric analysis	Step 8: Development of item bank
Who?	Program leaders	Workshop	Workshop	Workshop	Review panel	Remote delivery	SJT consultants	SJT consultants

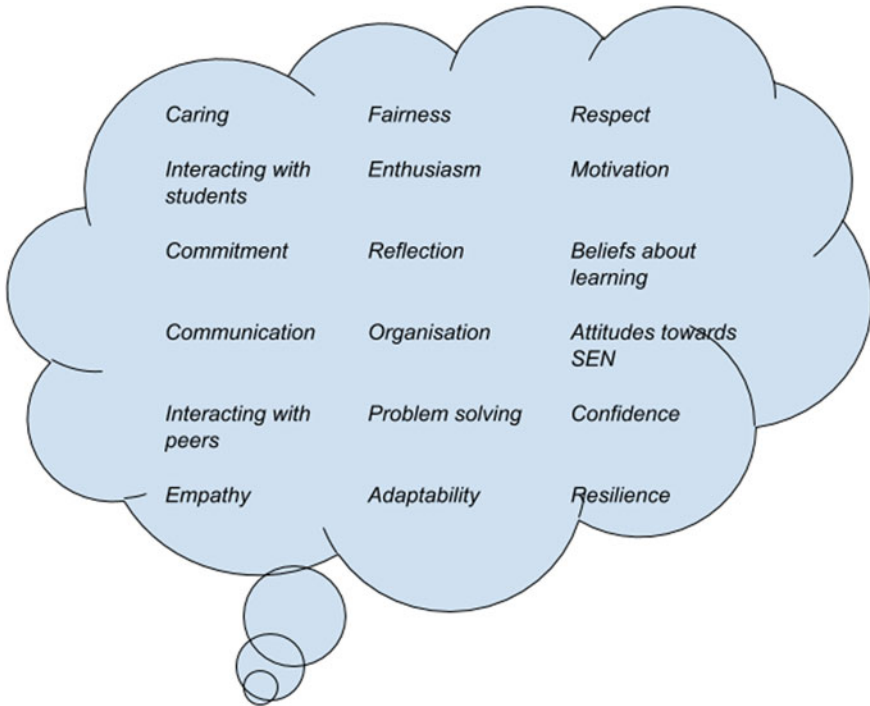


**Fig. 7.3** Proposed framework for selection of teachers for training and professional practice

storage, scoring, and general ease-of-use. The response format issue has been the subject of considerable research, with the overall finding that a rating format is preferable over ranking or ‘pick best, pick worst’ formats (Arthur et al., 2014).

**Step 2: Identify and define key attributes.** In most cases, the key purpose for implementing SJTs for teacher selection is to evaluate applicants’ non-cognitive attributes—the ‘soft’ skills that are so difficult to assess in a reliable and valid fashion at interview. SJTs can be designed to measure a range of attributes depending on how they are developed, but most researchers agree that SJTs measure an individual’s awareness and judgment about effective behavior in specific situations. Assessing applicant judgment in a fair and reliable way is the hallmark of SJTs, but which attributes are best targeted when constructing SJTs? An important development step is to identify a set of key attributes on which to build scenario content. A brainstorming session can be part of SJT development; in a workshop environment, the questions are *What are the key non-cognitive (or non-academic) attributes of novice teachers? What are the attributes that are necessary for the success of trainees? Are there particular attributes that are especially important in this context?* The results from the brainstorming session (see Fig. 7.4 below) are collated and assessed through a content analysis of collected data, with a frequency analysis indicating the agreed upon attributes on which to build scenarios. It is suggested that between 3–6 attributes or attribute clusters be targeted for scenario development.

**Step 3 (optional): Review of existing items in target context.** Using an ‘off-the-shelf’ SJT for selecting teachers offers certain advantages, such as proven psychometric properties, efficiencies of cost and time, and content that has been shown to evoke positive applicant reactions. However, adaptations to existing items may be necessary if the target context differs in significant ways from the context in which the test was originally developed. The level of adaptations depends on the degree of contextual differences: a review of scenarios, response options, and scoring will usually highlight any changes needed, which may range from terminology (replacing ‘headteacher’ for ‘principal’, for example when adapting a UK-developed SJT for use in Canada) to revisions of scenarios, response options, and scoring if the cultural distance is considerable. In the section ‘Adapting content for new settings’, we provide a more thorough description of the adaptation process when cultural distance is large.



**Fig. 7.4** Results from brainstorm session on key non-cognitive attributes for novice teachers

**Step 4: Generate new items using a ‘critical incidents’ approach.** An SJT item comprises a scenario that describes a realistic work-place situation, and a set of plausible options for responding to the scenario. Developing new items for teacher selection SJTs demands a knowledge of the relevant context, a knowledge of typical challenges facing new teachers, and the relevant experience needed to understand the pros and cons of various response options. Scenarios are typically built using a critical incidents approach (e.g., Buyse & Lievens, 2011) in which instructions are sent to workshop participants along the lines of ‘We are developing a teacher selection tool that focuses on the non-academic attributes associated with successful teaching (e.g., empathy, conscientiousness, adaptability). In our upcoming workshop, we will ask you to share scenarios of incidents that are related to these attributes. The scenarios should reflect situations that novice teachers encounter and should be related to one of the target attributes’. During the workshop, participants generate scenarios and response options that are subsequently reviewed by test developers (to eliminate errors, inappropriate and redundant items, and items that do not map onto the target attributes).

**Step 5. Conduct a review panel to set scoring.** The purpose of the review panel is to carefully scrutinize the scenarios and response options generated in Step 4, and to answer the questions *Are the items set in the correct context? Are the response options feasible and set at an appropriate level for a novice teacher? Does the content depend too heavily on specific procedural knowledge?* The reviewed and revised items are then tested in a concordance panel (which can be delivered remotely) in which experienced teachers complete the prototype SJT to determine the level of consensus of scoring of the SJT and to provide additional feedback on the items. Items with high consensus are retained for use in next steps; items with low consensus are revised and assessed in an iterative manner.

**Step 6. Design and pilot SJT.** Items that fare well in Step 5 are tabulated against the target attributes, and a representative selection of items are included in the pilot SJT and administered either to (a) ‘incumbents’ (i.e., students already admitted to ITE programs, or beginning teachers who are working in schools) or to (b) applicants for ITE programs. When pilot SJTs are delivered to applicants, the pilot tests are prefaced with a statement indicating that completion is voluntary and will not affect their application status. Applicant reaction data is importantly collected at this stage, and concurrent validity data (i.e., interview scores, academic data, teaching performance data if incumbents). After administration, scoring keys may be adjusted based on psychometric analysis in Step 7.

**Step 7: Psychometric analysis.** Once the data is collected, scoring keys are set using one of four approaches: rational, theoretical, empirical, or integrated. A rational approach uses an SME consensus approach as discussed in Step 5. A theoretical approach builds a scoring key based on what theory suggests is the ‘best’ course of action in a situation. An empirical approach is determined by evaluations of the relations between applicant responses and an external criterion (e.g., other interview scores, teaching ratings). An integrated approach sets the initial scoring key using a rational (or theoretical) approach, and then revises the key based on empirical results.

Analysis of SJT data typically includes a measure of reliability (internal consistency, test–retest, or split-half), item difficulty, and concurrent, construct, and/or predictive validity. Reliability estimates (consistency of measurement) are notoriously fickle for SJTs due to their multidimensional nature, and reliability indices other than conventional Cronbach’s alpha are recommended (e.g., test–retest, split-half using the Spearman-Brown prophecy formula), and if alpha is used, should be considered a lower bound of the reliability estimate (Whetzel et al., 2020).

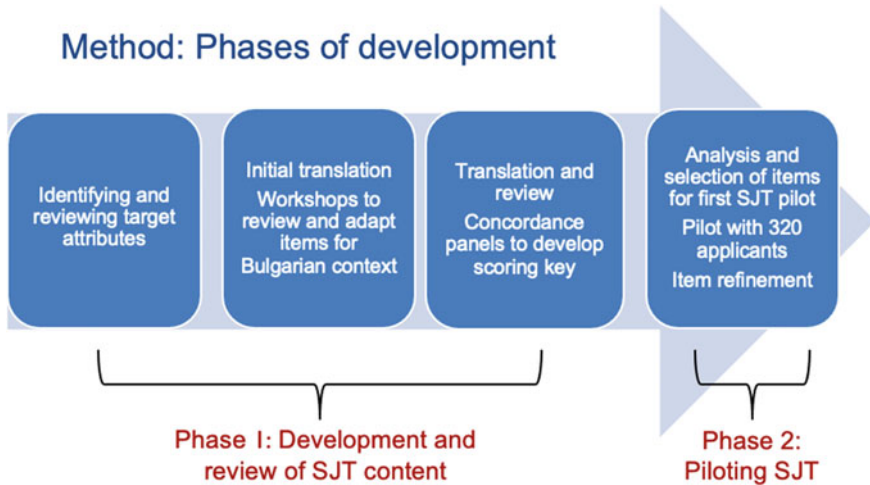
**Step 8: Development of item bank.** Items that are not included in the development of the SJT will be retained in an item bank, with some items rated as ‘good’ (items with scoring consensus and acceptable psychometric properties) that can be included in future test versions, and some items rated as ‘needing further work’ that can be revised and re-piloted or discarded. A functioning item bank is important to develop future iterations of the test and is important to bolster test security.

## 7.4 Adapting Content for New Settings

We saw in Chap. 3 how culture—the shared beliefs, goals, and values that guide the way we think and behave—influences the education environment and even influences the way the personal characteristics deemed necessary for novice teachers' success are viewed. For teacher selection methods, it is not a case of *one-size-fits-all* when it comes to using off-the-shelf selection methods, and although methods such as SJTs or MMIs that are effective in one setting may prove effective in other settings, the *content* of these methods needs to reflect the cultural, social, and educational context. In a study exploring the feasibility of using an American-developed SJT assessing integrity in a Spanish context, Lievens et al. (2015) found that most of the scenarios (84%) were deemed to be realistic by Spanish test-takers, with similar relations to external criteria. Herde et al. (2019) tested the measurement invariance of five SJTs testing 'twentieth century skills' (e.g., achieving objectives, adapting to change) across multiple countries in Europe and Latin America, and found the same latent factorial structure (and similar internal consistency coefficients) across regional groups, suggesting participants interpreted the SJT scenarios and response options in the same way. Nevertheless, merely translating SJT content into a new language is likely to be insufficient especially with high levels of 'cultural distance'; a deep knowledge of the target culture is needed to ensure that items reflect the social, cultural, and educational norms that may be different from those in the context where the test was first developed.

Especially in the case where cultural distance is great, SJT development requires socio-cultural awareness and collaboration with partners in the target setting. A true partnership is needed: developers of selection tests who are serving as consultants will have only moderate knowledge of the cultural setting; education experts in the target setting may have only moderate knowledge of test-writing. Ryan and Brunfaut (2016) used a case study approach to better understand how to conduct (language) test development work in cultural settings that are unfamiliar to the test developers. They offered several suggestions to maximize the chance of effective test-writing. First, the test developers benefit from preliminary work to increase familiarization with the target language and culture, ensuring a level of basic knowledge about the language and socio-cultural and educational context. In parallel, the education experts in the target culture benefit from preliminary work on familiarization with test-development principles. In the case of SJT writing, education experts can be provided with sample items, critical incident prompt materials, and a description of key psychometric considerations including item discrimination, reliability, item difficulty, and predictive validity.

Researchers at the Teacher Selection Project have developed a range of teacher selection tools across cultural contexts, not only in the UK, but in settings that are culturally distant from their original work in England (e.g., Bulgaria, Lithuania, Malawi, and Morocco). The process of adaptation began with a determination of appropriateness of key non-cognitive attributes. A cross-national comparison



**Fig. 7.5** Adapting SJTs for Bulgarian context

(Klassen et al., 2018) found that core non-cognitive attributes (i.e., communication, adaptability, organization) identified in England were endorsed across culturally disparate settings (i.e., Finland, Malawi, Oman), but each non-English setting also proposed additional non-cognitive attributes believed to be essential to successful novice teaching. In line with the 2018 study, some non-cognitive attributes of successful novice teachers seem to be universal, and some appear to be context- and culture-specific.

**Case study: Bulgaria.** An existing English-language SJT originally developed in the UK was adapted for use in Bulgaria (Rushby & Klassen, 2019). Figure 7.5 presents the two phases of SJT development, with Phase 1 involving the development and review of SJT content, and Phase 2 including the pilot-testing and revision of content. After reviewing and confirming target attributes established in previous work (i.e., empathy and communication, organization and planning, resilience and adaptability), the translated SJTs were adapted. The translation of existing items included surface changes to scenarios (school contexts, teaching roles) and deeper changes to item responses (desired and undesired options for responses).

A review panel to establish the scoring key that included 28 subject matter experts (SMEs) was conducted to review the existing scoring and to identify how expert teachers in Bulgaria evaluated scenarios and scoring originally developed in England. In cases where scenarios were not substantially changed (i.e., apart from names and job titles), most of the scoring (57.6%) of the responses was the same across contexts, about one-third of responses (28.6%) were one position away (e.g., ‘appropriate’ in the UK setting; ‘somewhat appropriate’ in Bulgaria), 6.2% of responses were two positions away, and one response was three positions away (e.g., ‘inappropriate’ in UK; ‘appropriate’ in Bulgaria). The scenario that showed the greatest cross-cultural

difference was a classroom situation where a teacher assistant was routinely interrupting and correcting a teacher's lessons: confronting the assistant in front of the students was deemed 'appropriate' by most Bulgarian SMEs, but 'inappropriate' by most UK SMEs. Identifying and discussing these differences in interpersonal relationships was an important part of the adaptation process in this context.

**Case study: rural and remote Australia.** Durksen and Klassen (2018) developed an SJT to promote the key characteristics needed for rural and remote Australian settings. In Australia, the turnover rate for teachers in rural and remote settings is up to six times higher than in city schools, with many new teachers in remote regions leaving their posts before the end of their contracted teaching (Lyons, 2006). In this project sponsored by the New South Wales Department of Education, the authors began by evaluating the key clusters of non-cognitive attributes that had been developed in the UK: resilience and adaptability, organization and planning, and empathy and communication. A review panel of experienced teachers concluded that the non-cognitive attributes previously identified in the UK were universally salient in a remote and rural setting, but that a new attribute cluster—'culture and context' was necessary to capture the unique qualities needed for success in the target culture. The new cluster was defined as 'The capability to adapt to remote settings, recognize the importance of building relationships and maintaining professional behavior in all aspects of community life. Demonstrates sensitivity to cultural knowledge and practice'. After identifying the key attributes needed for success in remote settings, an item-writing workshop resulted in 37 new NSW-specific items and some minor revisions to the existing 32 UK items trialed in earlier stages. Although the cultural distance between the UK and Australia was not as great as between the UK and Bulgaria, important cultural differences were identified, and SJT development needed to reflect these differences.

## 7.5 Chapter Summary

Current teacher selection tools are not always reflective of the most recent research in organizational or educational psychology, but recent studies have shown that SJTs provide an evidence-supported alternative for organizations looking to select effective teachers (Klassen & Kim, 2019). In this chapter we first explored the use of SJTs for selection in detail, and considered how SJTs could be built deductively, i.e., using a top-down, construct-driven approach, or inductively, i.e., using a bottom-up, inductive approach in collaboration with subject matter experts. A blueprint for developing SJTs was provided alongside a discussion on how to adapt SJTs to new contexts. In the next chapter we examine another research-supported method for selection—multiple mini-interviews, or MMIs—that are now being implemented and tested for teacher selection in the UK and Finland.



## References

- Abozaid, H., Park, Y. S., & Tekian, A. (2017). Peer review improves psychometric characteristics of multiple choice questions. *Medical Teacher, 39*, S50–S54.
- Arthur, W., Jr., Glaze, R. M., Jarrett, S. M., White, C. D., Schurig, I., & Taylor, J. E. (2014). Comparative evaluation of three situational judgment test response formats in terms of construct-related validity, subgroup differences, and susceptibility to response distortion. *Journal of Applied Psychology, 99*, 1–11.
- Bandura, A. (1999). A social cognitive theory of personality. In L. Pervin & O. John (Eds.), *Handbook of personality* (2nd ed., pp. 154–196). Guilford Publications.
- Bardach, L., Rushby, J. V., Kim, L. E., & Klassen, R. M. (2020). Using video- and text-based situational judgement tests for teacher selection: A quasi-experiment exploring the relations between test format, subgroup differences, and applicant reactions. *European Journal of Work and Organizational Psychology*. <https://doi.org/10.1080/1359432X.2020.1736619>
- Bateson, J. E., Wirtz, J., Burke, E. F., & Vaughan, C. J. (2014). Sifting to efficiently select the right service employees. *Organizational Dynamics, 43*, 312–320.
- Bowles, T., Hattie, J., Dinham, S., Scull, J., & Clinton, J. (2014). Proposing a comprehensive model for identifying teaching candidates. *The Australian Educational Researcher, 41*, 365–380.
- Boyatzis, R. E., & Kelner, S. P. (2010). Competencies as behavioral manifestation of implicit motives. In O. C. Schultheiss & J. C. Brunstein (Eds.), *Implicit motives* (pp. 488–509). Oxford University Press.
- Bruk-Lee, V., Lanz, J., Drew, E. N., Coughlin, C., Levine, P., Tuzinski, K., & Wrenn, K. (2016). Examining applicant reactions to different media types in character-based simulations for employee selection. *International Journal of Selection and Assessment, 24*, 77–91.
- Buyse, T., & Lievens, F. (2011). Situational judgment tests as a new tool for dental student selection. *Journal of Dental Education, 75*, 743–749.
- Carter Andrews, D. J., Castro, E., Cho, C. L., Petchauer, E., Richmond, G., & Floden, R. (2019). Changing the narrative on diversifying the teaching workforce: A look at historical and contemporary factors that inform recruitment and retention of teachers of color. *Journal of Teacher Education, 70*, 6–12.
- Catano, V. M., Brochu, A., & Lamerson, C. D. (2012). Assessing the reliability of situational judgment tests used in high-stakes situations. *International Journal of Selection and Assessment, 20*, 334–346.
- Christian, M. S., Edwards, B. D., & Bradley, J. C. (2010). Situational judgment tests: Constructs assessed and a meta-analysis of their criterion-related validities. *Personnel Psychology, 63*, 83–117. <https://doi.org/10.1111/j.1744-6570.2009.01163.x>
- Clevenger, J., Pereira, G. M., Wiechmann, D., Schmitt, N., & Harvey, V. (2001). Incremental validity of situational judgment tests. *Journal of Applied Psychology, 86*, 410–417. <http://dx.doi.org/https://doi.org/10.1037/0021-9010.86.3.410>
- Durksen, T. L., & Klassen, R. M. (2018). The development of a situational judgment test (SJT) of personal attributes for quality teaching in rural and remote Australia. *The Australian Educational Researcher, 45*, 255–276.
- Elliott, J. G., Stemler, S. E., Sternberg, R. J., Grigorenko, E. L., & Hoffman, N. (2011). The socially skilled teacher and the development of tacit knowledge. *British Educational Research Journal, 37*, 88–103. <https://doi.org/10.1080/01411920903420016>
- Fan, J., Stuhlman, M., Chen, L., & Weng, Q. (2016). Both general domain knowledge and situation assessment are needed to better understand how SJTs work. *Industrial and Organizational Psychology, 9*, 43–47.
- Freudenstein, J.P., Schäpers, P., Roemer, L., Mussel, P., & Krumm, S. (2020). Is it all in the eye of the beholder? The importance of situation construal for situational judgment test performance. *Personnel Psychology*, DOI: <https://doi.org/10.1111/peps.12385>.

- Harris, A. M., Siedor, L. E., & Listyg, & Carter, N. T. . (2016). In defense of the situation: An interactionist explanation for performance on situational judgment tests. *Industrial and Organizational Psychology*, 9, 23–28.
- Herde, C. N., Lievens, F., Solberg, E. G., Harbaugh, J. L., Strong, M. H., & Burkholder, G. J. (2019). Situational judgment tests as measures of 21<sup>st</sup> century skills: Evidence across Europe and Latin America. *Journal of Work and Organizational Psychology*, 35, 65–74.
- Hooper, A. C., Jackson, H. L., & Motowidlo, S. J. (2004). *Situational judgment measures of personality and work-relevant performance*. Paper presented at the 112th annual meeting of the American Psychological Association, Honolulu, HI.
- Kasten, N., & Freund, P. A. (2016). A meta-analytical multilevel reliability generalization of situational judgment tests (SJTs). *European Journal of Psychological Assessment*, 32, 230–240.
- Klassen, R. M., Durksen, T. L., Al Hashmi, W., Kim, L. E., Longden, K., Metsäpelto, R.-L., Poikkeus, A. M., & Györi, J. (2018). Cultural context and teacher characteristics: Exploring the non-cognitive attributes of prospective teachers in four countries. *Teaching and Teacher Education*, 72, 64–74.
- Klassen, R. M., Durksen, T. L., Kim, L. E., Patterson, F., Rowett, E., Warwick, J., Warwick, P., & Wolpert, M. A. (2017). Developing a proof-of-concept selection test for entry into primary teacher education programs. *International Journal of Assessment Tools in Education*, 4, 96–114.
- Klassen, R. M., Durksen, T. L., Rowett, E., & Patterson, F. (2014). Applicant reactions to a situational judgment test used for selection into initial teacher training. *International Journal of Educational Psychology*, 3, 104–125. <https://doi.org/10.4471/ijep.2014.07>
- Klassen, R. M., & Kim, L. E. (2019). Selecting teachers and prospective teachers: A meta-analysis. *Educational Research Review*, 26, 32–51.
- Klassen, R. M., Kim, L. E., Rushby, J., & Bardach, L. (2020). Can we improve how we screen applicants for initial teacher education? *Teaching and Teacher Education*, 87, 102949.
- Klassen, R. M., & Rushby, J. V. (2019, January). Can we predict preservice teachers' performance in teaching placements? The validity of ITE selection methods. *TSP Working Paper*. <https://bit.ly/2SeQkVv>
- Koczwara, A., Patterson, F., Zibarras, L., Kerrin, M., Irish, B., & Wilkinson, M. (2012). Evaluating cognitive ability, knowledge tests and situational judgement tests for postgraduate selection. *Medical Education*, 46(4), 399–408.
- Lievens, F., & De Soete, B. (2015). Situational judgment tests. In J. D. Wright (Ed.), *International Encyclopedia of the Social & Behavioral Sciences* (2nd ed., Vol. 22, pp. 13–19). Elsevier.
- Lievens, F., & Motowidlo, S. J. (2016). Situational judgment tests: From measures of situational judgment to measures of general domain knowledge. *Industrial and Organizational Psychology*, 9, 3–22.
- Lievens, F., & Sackett, P. R. (2006). Video-based versus written situational judgment tests: A comparison in terms of predictive validity. *Journal of Applied Psychology*, 91, 1181–1188.
- Lievens, F., & Sackett, P. R. (2012). The validity of interpersonal skills assessment via situational judgment tests for predicting academic success and job performance. *Journal of Applied Psychology*, 97, 460–468.
- Lievens, F., Corstjens, J., Sorrel, M. A., Abad, F. J., Olea, J., & Ponsoda, V. (2015). The cross-cultural transportability of situational judgment tests: How does a US-based integrity situational judgment test fare in Spain? *International Journal of Selection and Assessment*, 23, 361–372.
- Lyons, T. (2006). Science, ICT and mathematics education in rural and regional Australia, *SIMERR National Research Centre [Report]*. <http://apo.org.au/node/2732>.
- Patterson, F., Ashworth, V., Mehra, S., & Falcon, H. (2012). Could situational judgement tests be used for selection into dental foundation training? *British Dental Journal*, 213, 23–26.
- Patterson, F., Lievens, F., Kerrin, M., Munro, N., & Irish, B. (2013). The predictive validity of selection for entry into postgraduate training in general practice: evidence from three longitudinal studies. *British Journal of General Practice*, November, e734–e741. doi: <https://doi.org/10.3399/bjgp13X674413>

- Patterson, F., Tavabie, A., Denney, M., Kerrin, M., Ashworth, V., Koczwara, A., & MacLeod, S. (2013). A new competency model for general practice. *British Journal of General Practice*, May, e331–e338. doi: <https://doi.org/10.3399/bjgp13X667196>
- Patterson, F., Zibarras, L., & Ashworth, V. (2015). Situational judgement tests in medical education and training: Research, theory and practice: AMEE Guide No. 100. *Medical Teachers*, Aug 27, 1–15 E-pub ahead of print.
- Patterson, F., Zibarras, L., Carr, V., Irish, B., & Gregory, S. (2011). Evaluating candidate reactions to selection practices using organisational justice theory. *Medical Education*, 45, 289–297.
- Rushby, J. V., & Klassen, R. M. (2019, September). *The cross-cultural transferability of a situational judgment test used for initial teacher selection*. Paper presented at the British Psychological Society Psychology of Education Section annual meeting: Manchester, UK.
- Ryan, E., & Brunfaut, T. (2016). When the test developer does not speak the target language: The use of language informants in the test development process. *Language Assessment Quarterly*, 13, 393–408.
- Shultz, M. M., & Zedeck, S. (2012). Admission to law school: New measures. *Educational Psychologist*, 47, 51–65. <https://doi.org/10.1080/00461520.2011.610679>
- Tiffin, P. A., Paton, L. W., O'Mara, D., MacCann, C., Lang, J. W., & Lievens, F. (2020). Situational judgement tests for selection: Traditional vs construct-driven approaches. *Medical Education*, 54(2), 105–115.
- Whetzel, D. L., & McDaniel, M. A. (2009). Situational judgment tests: An overview of current research. *Human Resource Management Review*, 19, 188–202. <https://doi.org/10.1016/j.hrmr.2009.03.007>
- Whetzel, D. L., Sullivan, T. S., & McCloy, R. A. (2020). Situational judgment tests: An overview of development practices and psychometric characteristics. *Personnel Assessment and Decisions*, 6, 1–16.